

# Person Detection in Collaborative Group Learning Environments Using Multiple Representations

Wenjing Shi<sup>1</sup>, Marios S. Pattichis<sup>1</sup>, Sylvia Celedón-Pattichis<sup>2</sup> and Carlos LópezLeiva<sup>2</sup>

<sup>1</sup> *image and video Processing and Communications Lab (ivpcl.unm.edu)*

*Dept. of Electrical and Computer Engineering*

University of New Mexico, United States.

<sup>2</sup> *Dept. of Language, Literacy, and Sociocultural Studies*

University of New Mexico, United States.

{wshi, pattichi, sceledon, callopez}@unm.edu

**Abstract**—We introduce the problem of detecting a group of students from classroom videos. The problem requires the detection of students from different angles and the separation of the group from other groups in long videos (one to one and a half hours).

We use multiple image representations to solve the problem. We use FM components to separate each group from background groups, AM-FM components for detecting the back-of-the-head, and YOLO for face detection. We use classroom videos from four different groups to validate our approach. Our use of multiple representations is shown to be significantly more accurate than the use of YOLO alone.

**Index Terms**—person detection, video analysis, AM-FM representations.

## I. INTRODUCTION

We study the problem of person detection in collaborative learning environments' videos. The problem has some unique challenges associated with detecting students sitting around a table.

We present an example of the collaborative learning environment in Figure 1. The students' faces are imaged from different angles. They are at different distances from the camera. In many cases, the faces are not visible. Furthermore, there are other groups in the background. For the purposes of this paper, we are only interested in detecting students that are sitting around the table that is closest to the camera. All other students are not to be included in the analysis.

Currently, human detection methods are dominated by neural network methods. As an example, in [1], the authors used a lightweight Convolutional Neural Network (L-CNN) to detect humans in surveillance video frames. In another example, in [2], the authors used a multi-stream multitask deep network for joint human detection and head poses estimation in RGB-D videos.

We also provide a summary of prior research on classroom videos. In [3], we considered using K-NN classifiers with AM-FM representations for person detection. In [4], the combination of color and FM representations was considered for face detection. In [4], back-of-the-head detection was performed using AM-FM representations. In [5], the method in [4] was

extended to detect where the students were looking. The importance of FM representations for face detection was further documented in [6]. In [7], the authors used head detection to detect talking activities. In [8], the authors developed methods for hand detection. In [9], the authors considered the use of YOLO [10] for head detection to build a bilingual speech recognition system. Fast video face detection was recently described in [11].

The current paper extends prior methods through the combination of YOLO with AM-FM representations. Firstly, YOLO is used to process RGB images for face detection. Secondly, FM images, characterized by higher instantaneous frequencies, are used with LeNet5 to remove non-group faces that were falsely detected by YOLO. Thirdly, LeNet5 is used to remove false positives from the back-of-the-head classifier.

We summarize the rest of the paper in three additional sections. The proposed method is summarized in section II. The results are given in section III. Concluding remarks are given in section IV.

## II. METHOD

We present a system diagram of the entire system in Figure 2. YOLO V3 is used for face detection. The detected



Fig. 1: An example of a group of students participating in a collaborative learning environment.

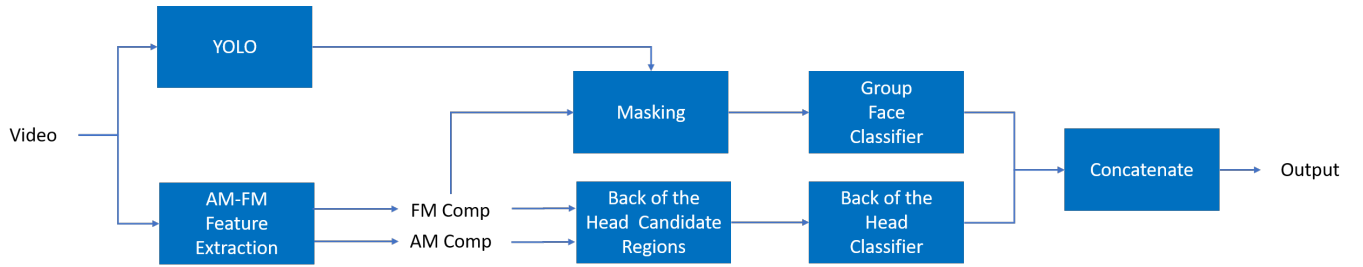


Fig. 2: Student Group Detection System.



Fig. 3: AM-FM representation for classroom environment. (a) Classroom image. (b) AM component. (c) FM component

faces are further processed in combination with the AM-FM components as described below.

AM-FM components are extracted from the grayscale (Y-component) using dominant component analysis (DCA) estimated using a 54-channel Gabor filterbank as described in [4]. Using DCA, the input image frame is approximated by:  $I(x, y) \approx a(x, y) \cos \varphi(x, y)$  where  $a(x, y)$  denotes the AM component and  $\cos \varphi(x, y)$  denotes the FM component. Figure 3 shows the extracted AM-FM components.

The FM image is masked by the results of the YOLO face detector. We apply this step to extract the FM components over students within the desired group as well as other groups. FM components over the faces of the closest group will exhibit lower frequency components than the higher frequency components associated with distant faces from other groups. To detect the group faces, we thus apply a simple, LeNet-based classifier [12] on the extracted FM components over  $100 \times 100$  pixel regions.

The AM-FM components are also used to detect the hair and back-of-the-head candidate regions described in [4]. A LeNet based classifier is used to detect the back-of-the-heads against background detections as detailed in [4]. For each video frame, we detect the entire group by concatenating the results from the face and back-of-the-head classifiers.

### III. EXPERIMENTAL DATA AND RESULTS

The proposed methodology was trained and tested on digital videos recorded through actual classroom implementations of the Advancing Out-of-school Learning in Mathematics and Engineering (AOLME) program. The videos depicted a variety of different learning environments with rich background activities and several background groups.

For training the YOLO face detector, we used 1000 faces and 1200 non-face images from student groups extracted from

54 different videos. Among the selected face images, we used 70% of the images for training and 30% for validation.

For training the group face classifier, refer to Table I. The dataset was generated from the same 54 videos from 13 different groups. As summarized in Table I, the augmented dataset contained about 70,000 group face images and 70,000 non-group face images.

TABLE I: Group faces classifier training, validation, and testing. The numbers include seven-fold data augmentation performed using random rescaling, cropping, rotating, and flipping.

	Group Faces	Non-group Faces
<b>Training</b>	39,232	39,259
<b>Validation</b>	16,813	16,825
<b>Testing</b>	14,011	14,021
<b>Total</b>	<b>70,056</b>	<b>70,105</b>
<b>AUC Score</b>	0.97	
<b>Accuracy</b>	97.5%	

TABLE II: Back-of-the-head classifier training and testing. The numbers include seven-fold data augmentation performed using random rescaling, cropping, rotating, and flipping.

	Back-of-the-Heads	Other
<b>Training+Validation</b>	22,768	22,800
<b>Testing</b>	5,710	5,682
<b>Total</b>	<b>28,478</b>	<b>28,482</b>
<b>AUC Score</b>	0.97	
<b>Accuracy</b>	97.3%	

Table II describes the training and validation dataset for the back-of-the-head classifier. The dataset uses 56,000 frames

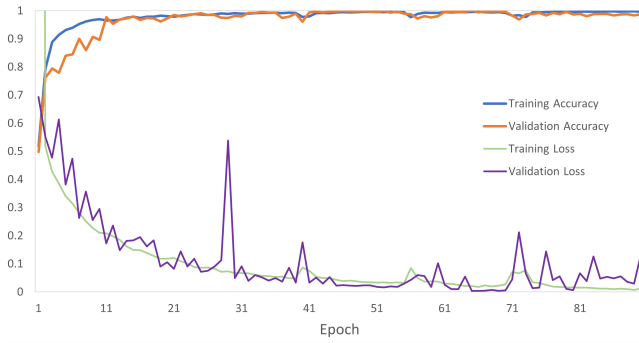


Fig. 4: Group faces classifier training.

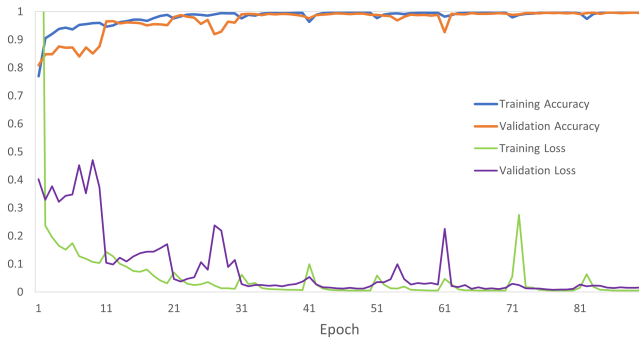


Fig. 5: Back-of-the-head classifier training.

from the 54 videos. The dataset was used to train a second LeNet5 classifier to remove false positives when detecting the back-of-the-head regions.

For the two LeNet5 classifiers, we allocated 70% for training and 30% for validation. An independent dataset, 20% of the total images, was used for testing the final model in each case. The training datasets include a seven-fold data augmentation performed using random rescaling, cropping, rotating, and flipping. The training and validation accuracies are provided in Figures 4 and 5. According to the table, we get over 97% AUC score and accuracy for each model.

We used a new set of four long videos from different student groups for the final testing. The video results are given in Table III. For successful detection, we require the intersection over union (IOU) score to be at least 0.6. From the results, it is clear that the proposed approach outperformed the use of YOLO V3 alone.

We show examples of true positives, false positives, and false negatives in Fig. 6. False positives are associated with out-of-group detection. False negatives are due to occlusion.

Figure 7 displays the comparative examples, YOLO V3 only, ground truth, and our proposed method. From the results, it is clear that YOLO V3 cannot differentiate between the in-group and out-of-group faces. The use of the AM-FM components allows us to remove the out-of-group faces, as shown in the right column of Figure 7.

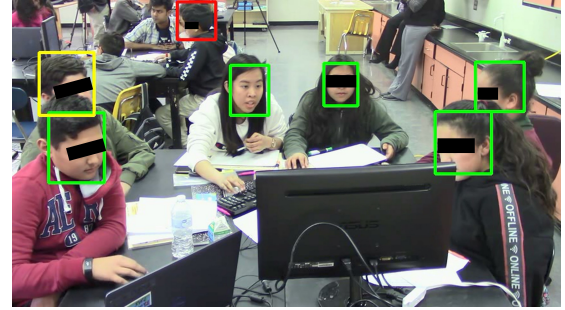
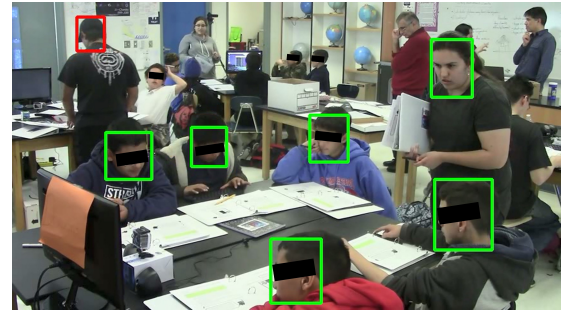


Fig. 6: Head detection system results. True positives are bounded by green boxes. False positives are bounded by red boxes. False negatives are bounded by yellow boxes.

#### IV. CONCLUSION

The paper presents a method for detecting groups of students using multiple image representations. The effective combination of YOLO V3 with AM-FM representations provides for improved results. Our current research is focused on face recognition and talking activity detection.

#### ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1613637, No.1842220 and, No.1949230.

#### REFERENCES

- [1] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan, "Real-time human detection as an edge service enabled by a lightweight cnn," in *2018 IEEE International Conference on Edge Computing (EDGE)*. IEEE, 2018, pp. 125–129.
- [2] G. Zhang, J. Liu, H. Li, Y. Q. Chen, and L. S. Davis, "Joint human detection and head pose estimation via multistream networks for rgb-d videos," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1666–1670, 2017.
- [3] W. Shi, "Human Attention Detection Using AM-FM Representations," Master's thesis, the University of New Mexico, Albuquerque, New Mexico, 2016.
- [4] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Robust head detection in collaborative learning environments using am-fm representations," in *2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*. IEEE, 2018, pp. 1–4.
- [5] —, "Dynamic group interactions in collaborative learning videos," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 1528–1531.
- [6] L. S. Tapia, M. S. Pattichis, S. Celedón-Pattichis, and C. L. Leiva, "The importance of the instantaneous phase for face detection using simple convolutional neural networks," in *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*. IEEE, 2020, pp. 1–4.

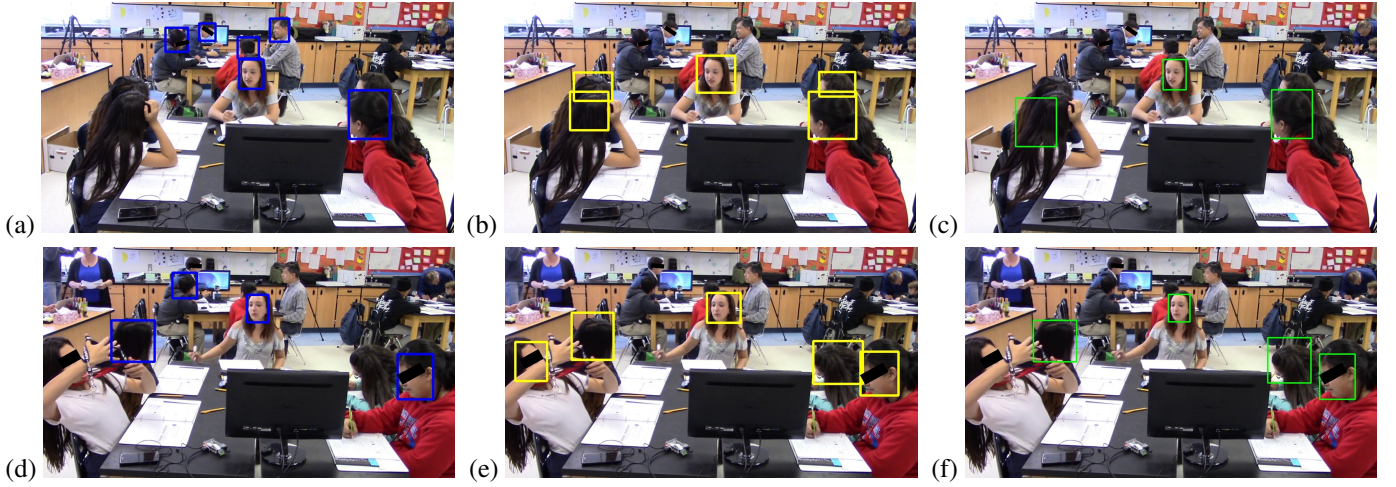


Fig. 7: Examples from Group Detection Results. Left column ((a) and (d)) shows the results of YOLO V3. Middle-column ((b) and (e)) shows the ground truth. Right-column ((c) and (f)) shows the results of the proposed method.

TABLE III: Comparative results for student group detection over four videos. TP, FP, FN refer to true positives, false positives, and false negatives, respectively. F1 scores are given for each video and each method. The videos represent different student groups.

Video	Length in minutes	Labeled Students	Method	Detected Students	TP	FP	FN	F1
V1	96 minutes	1,627,320	YOLO	1,915,935	1,153,959	761,976	124,527	0.72
			Proposed Method	1,397,790	1,183,630	214,160	344,640	<b>0.81</b>
V2	85 minutes	887,700	YOLO	1,274,429	723,283	551,146	12,153	0.72
			Proposed Method	847,250	728,110	119,140	110,140	<b>0.86</b>
V3	117 minutes	1,063,300	YOLO	792,291	720,762	715,29	321,159	0.79
			Proposed Method	819,700	745,880	73,820	293,640	<b>0.80</b>
V4	108 minutes	1,139,850	YOLO	1,212,963	839,450	373,513	120,252	0.77
			Proposed Method	950,210	859,290	90,920	242,850	<b>0.84</b>

- [7] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Talking detection in collaborative learning environments," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2021, pp. 242–251.
- [8] S. Teeparthi, V. Jatla, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Fast hand detection in collaborative learning environments," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2021, pp. 445–454.
- [9] L. Sanchez Tapia, A. Gomez, M. Esparza, V. Jatla, M. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Bilingual speech recognition by estimating speaker geometry from video data," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2021, pp. 79–89.
- [10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [11] P. Tran, M. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Facial recognition in collaborative learning videos," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2021, pp. 252–261.
- [12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.